Writing Latinized Taiwanese Languages with Unicode, 1999 Edition

TÈ Khái-sū*
Taiwan Protocol

January 1, 1999

1 Overview

The trend to use Unicode [1] as the universal character encoding for information interchange has been growing stronger, and most software producers on the market have announced plans to adopt Unicode. I have therefore decided to propose the appropriate way to encode information in Latinized Taiwanese languages with Unicode.

The Latinized Taiwanese languages, are ’Amis, Bunun, Hak-kâ-fa, Hō-ló-oē, Mandarin, Paiwan, Puyuma, Rukai, Saisiat, Tao, Tayal, Thao, Truku, and Tsou. Except for the case of Mandarin, where Hànyǔ Pīnyīn is considered, the Latinized orthographies I will be discussing are those found in the published Bible translations. Additional information found in other published materials using the same or only slightly different Latinized orthographies are also considered.

The most important parts of this article are sections 3, 4, 6 and 7, where characters not encoded in ISO 646 (ASCII) are used. Only such characters will be discussed in these sections.

2 ’Amis, Tao, Tayal and Truku

The Latinized forms of the languages ’Amis, Tao, Tayal and Truku are representable by the characters encoded by ISO 646 (ASCII). For the record, brief descriptions of the references are provided.

2.1 ’Amis

The new ’Amis translation of the Bible is still being typeset (in TEX), but I have a descriptive introduction to the Latinization [2] that most likely will be used by this new translation. According to this introduction, all characters used are in ISO 646.

*Also known as Kai-hsu Tai; email khaisu@formosa.org

2.2 Tao

All the characters in the Tao Bible [3] are ISO 646 characters.

2.3 Tayal

The Bible Society has told me that the Tayal Bible will soon be published. From a different source I have obtained an introductory description of the Latinization [4] published by the Presbyterian Church in Taiwan (Tâi-oân Ki-tok Tiú^{n}-ló Kàu-hoe̍). This Latinization is most likely that used in the Bible. All characters used are ISO 646 characters.

2.4 Truku

All characters used in the Truku Bible [5] are ISO 646 characters.

3 Bunun

The introduction to the Bunun Latinization [6] published by the Presbyterian Church and the Bible Society does not contain any non-ISO 646 characters, but the Hymnal [7] published by the Presbyterian Church contains two non-ISO 646 characters.

Đ U+0110 LATIN CAPITAL LETTER D WITH STROKE

đ U+0111 LATIN SMALL LETTER D WITH STROKE

The book has poor typesetting with all the strokes drawn with hand. The stroke on the small letter goes through the enclosure rather than then ascender, unlike the one shown here and in [1].

4 Hak-kâ-fa and Hō-ló-oē

Hak-kâ-fa and Hō-ló-oē share the characters composed by one of the Latin letters “a”, “e”, “i”, “o”, “u”, “m”, and “n” with one of the combining diacritics acute, grave, circumflex, macron (Hō-ló-oē only) and vertical line above. Relevant Unicode characters are listed below.

A	U+0041	LATIN CAPITAL LETTER A
E	U+0045	LATIN CAPITAL LETTER E
I	U+0049	LATIN CAPITAL LETTER I
M	U+004D	LATIN CAPITAL LETTER M
N	U+004E	LATIN CAPITAL LETTER N
O	U+004F	LATIN CAPITAL LETTER O
U	U+0055	LATIN CAPITAL LETTER U
a	U+0061	LATIN CAPITAL LETTER A
e	U+0065	LATIN CAPITAL LETTER E
i	U+0069	LATIN CAPITAL LETTER I
m	U+006D	LATIN CAPITAL LETTER M
n	U+006E	LATIN CAPITAL LETTER N
o	U+006F	LATIN CAPITAL LETTER O
u	U+0075	LATIN CAPITAL LETTER U
◌̀	U+0300	COMBINING GRAVE ACCENT
◌́	U+0301	COMBINING ACUTE ACCENT
◌̂	U+0302	COMBINING CIRCUMFLEX ACCENT
◌̄	U+0304	COMBINING MACRON
◌̍	U+030D	COMBINING VERTICAL LINE ABOVE

Unicode also has some precomposed characters used in both Hak-kâ-fa and Hō-ló-oē. They are listed below, not individually, but only in general categories with code ranges in order to save space. For an exhaustive, detailed list, see [8].

A, E, I, O, U, a, e, i, o *or* u + ◌̀, ◌́ *or* ◌̂	*in the range* U+00C0 → U+00FB
A, E, I, O, U, a, e, i, o *or* u + ◌̄	*in the range* U+0100 → U+016B
Ḿ, ḿ	U+1E3E, U+1E3F
Ń, ń	U+0143, U+0144

4.1 Hak-kâ-fa

In addition to the characters listed above, the Hakka Bible [9] of the Taiwanese Si-yen dialect uses one other vowel, "ṳ", and its combinations with acute, grave, circumflex, and vertical line above.

◌̤	U+0324	COMBINING DIAERESIS BELOW
Ṳ	U+1E72	LATIN CAPITAL LETTER U WITH DIAERESIS BELOW
ṳ	U+1E73	LATIN SMALL LETTER U WITH DIAERESIS BELOW

Phàng Tet̍-siû tries to present the pronunciations in both major Taiwanese Hak-kâ dialects (Si-yen and Hói-liuk̍) with the same Latinization in his introduction [10] and his dictionary [11]. The most important character is the Hói-liuk̍ seventh tone, denoted by a COMBINING RING ABOVE.

◌̊	U+030A					COMBINING RING ABOVE
Å	U+00C5	≡	U+0041	+	U+0324	LATIN CAPITAL LETTER A WITH RING ABOVE
Ů	U+016E	≡	U+0055	+	U+0324	LATIN CAPITAL LETTER U WITH RING ABOVE
å	U+00E5	≡	U+0061	+	U+0324	LATIN SMALL LETTER A WITH RING ABOVE
ů	U+016F	≡	U+0075	+	U+0324	LATIN SMALL LETTER U WITH RING ABOVE

Other special character used by Phàng are listed below.

o̲	U+0332	COMBINING LOW LINE
ọ	U+0323	COMBINING DOT BELOW
o̥	U+0325	COMBINING RING BELOW

4.2 Hō-ló-oē

In addition to the characters shared with Hak-kâ-fa, Hō-ló-oē have 3 characters: capital and small "o·", and "ⁿ"

O·	U+004F + U+00B7	
o·	U+006F + U+00B7	
ⁿ	U+207F	SUPERSCRIPT LATIN SMALL LETTER N

Typographically, the U+00B7 MIDDLE DOT should be properly raised, as it is here, to give the traditional appearance; some people may also prefer to have it kerned slightly.

A superscript capital N, "ᴺ", was used in all capital case in the dictionary of 1913 [12] and the Bible of 1933 [13][1], but this was not found in published materials in the second half of the 20th century, including [14] and [15][2].

The Bible of 1933 [13] also contained "wide diacritics" over both characters of the vowel "ng", for example, n͡g[3]. All other published materials I acquired presented this as "n̂g".

5 Mandarin

Since Unicode was designed with Hànyǔ Pīnyīn of Mandarin in mind, the presentation thereof could be easily deduced[4]; thus a discussion here would be redundant.

6 Paiwan

The Paiwan Bible [16] and the introduction to Paiwan Latinization [17] contained the following non-ISO 646 characters, all being Unicode characters.

o̱	U+0331			COMBINING MACRON BELOW
D̲	U+1E0E	≡ U+0044	+ U+0331	LATIN CAPITAL LETTER D WITH LINE BELOW
d̲	U+1E0F	≡ U+0064	+ U+0331	LATIN SMALL LETTER D WITH LINE BELOW
L̲	U+1E38	≡ U+004C	+ U+0331	LATIN CAPITAL LETTER L WITH LINE BELOW
l̲	U+1E39	≡ U+006C	+ U+0331	LATIN SMALL LETTER L WITH LINE BELOW
R̲	U+1E5E	≡ U+0052	+ U+0331	LATIN CAPITAL LETTER R WITH LINE BELOW
r̲	U+1E5F	≡ U+0072	+ U+0331	LATIN SMALL LETTER R WITH LINE BELOW
T̲	U+1E6E	≡ U+0054	+ U+0331	LATIN CAPITAL LETTER T WITH LINE BELOW
t̲	U+1E6F	≡ U+0074	+ U+0331	LATIN SMALL LETTER T WITH LINE BELOW

[1] For example, "JI̍-SÌᴺ" [12, p. 93], and "SAT-BÓ·-JÍᴺ" [13, p. 305].
[2] For example, "KO-IÛⁿ" [14, p. 1201], and "KÎ-THAⁿ" [15, Bo̍k-lio̍k].
[3] "mn̂g-chêng", [13, II Lia̍t-ông-kì 23:8, p. 444].
[4] See, for example, [1, pp. 7-26–7-27].

7 Tsou

No Bible translation is present for the Tsou language. I have obtained a Mandarin translation of the authoritative description of the Tsou language [18], and other books in the Tsou language by two Tsou experts Pǔ Zhōngyǒng and Pǔ Zhōngchéng. These books gave a Latinization used by the Christian church which used the following non-ISO 646 characters.

Ï	U+00CF	LATIN CAPITAL LETTER I WITH DIAERESIS
ï	U+00EF	LATIN SMALL LETTER I WITH DIAERESIS
Ʊ	U+0055 + U+0336	
ʉ	U+0289	LATIN SMALL LETTER U BAR

Although the capital letter u bar "Ʊ" is not found in the literature cited, it is conceptually necessary, and is composed using the capital letter "U" and U+0336 COMBINING LONG STROKE OVERLAY.

8 Puyuma, Rukai, Saisiat and Thao

I have obtained very little to no information on the following four languages: Puyuma, Rukai, Saisiat, and Thao, mostly due to the lack of published materials and the absence of Bible translations in these languages, which could in turn be accounted by their extremely small population of native speakers. My contact at the Bible Society told me that the Bible in Rukai will be published soon, but there still has been no telling about the other three languages. Because of the absence of applicable information, I am not going to further discuss these four languages in this article.

9 Acknowledgments

This article is not normative in any way, but implementation according to the proposal herein is encouraged. Suggestions, comments and corrections are more than welcome.

This article is dedicated to the people who are going to make it painless and elegant to typeset and transmit materials in Latinized Taiwanese languages.

Gratitude goes to the following people for their advices which contributed to this article: Harald T. ALVESTRAND, Michael EVERSON, Reni HACK, KÁN Chèng-gī, Jörg KNAPPEN, KÌ Goân-hong, LÂU Kiat-gák, Werner LEMBERG, LÎM Âng-koân, Rick MCGOWAN, Paul MCLEAN, ŃG Chong-gák, SHAN Chung-chieh, SO· Chi-bêng, TÂN Bêng-jîn, TĒⁿ Liông-kng, TĒⁿ Liông-úi, and Kenneth WHISTLER.

Supplemental information to this article can be obtained from the Taiwan Protocol web site at http://www.taiwanese.com/tp/.

References

[1] The Unicode Consortium. *The Unicode Standard, Version 2.0*. Addison-Wesley, Reading, Massachusetts, USA, 1996-07. ISBN 0-201-48345-9.

[2] Katayalan to Tilid no 'Amis. *Fa'lohay a Pintatanaman to Tilid no 'Amis*. Katayalan to Tilid no 'Amis, undated, published after 1964.

[3] Bible Society. *Seysyo No Tao (Yami New Testament)*. Bible Society, Taipei, Taiwan, 1994-11. Cat. No. YAMINT 263P, ISBN 957-99771-0-0.

[4] Duōào Yóugěihǎi and Ādòng Yóupàsī. *Lpgan Ke'na Tayal*. Tâi-oân Ki-tok Tiúⁿ-ló Kàu-hoē Tayal Tiong-hoē Bú-gú Thui-hêng Úi-oân-hoē, Hsinchu, Taiwan, 1991-08-01.

[5] Bible Society. *Soyang Patas (Holy Bible, Taroko and Today's Chinese)*. Bible Society, Taipei, Taiwan, 1988. Cat. No. TRTC73DIT.

[6] Tek-lâi Sín, editor. *malas bunun tu is-sisipul*. Bible Society, Taipei, Taiwan, 1987-10.

[7] Tâi-oân Ki-tok Tiúⁿ-ló Kàu-hoē Bunun Sèng-si Úi-oân-hoē, editor. *Chàn-bí-si*. Tâi-oân Ki-tok Tiúⁿ-ló Kàu-hoē Bunun Sèng-si Úi-oân-hoē, Taipei, Taiwan, 1992-04-28.

[8] Khái-sū Tè. Tâi-oân Hak-fa kap Hō-ló-oē sớ iōng ê jī-goân kap in ê Unicode hō-bé (characters used in Taiwanese Hak-fa and Hō-ló-oē and their Unicode encodings). Technical report, Taiwan Protocol, 1996-09-14. `http://www.taiwanese.com/tp/unirpt.txt`.

[9] Bible Society. *Hak-ngî Sṳn-kîn, Sîn-yuk lâu Sṳ-phiên: Hien-thoi Thòi-vân Hak-ngî Yit-pún (Hakka Bible, New Testament and Psalms: Today's Taiwan Hakka Version)*. Bible Society, Taipei, Taiwan, 1993. Cat. No. TTHV 363DI.

[10] Tet̍-siû Phàng. *Thai-kâ Lòi Hoȯk Hak-fa*. Nàm-thiên, Taipei, Taiwan, 1994-06. ISBN 957-638-017-0.

[11] Tet̍-siû Phàng. *Hak-kâ-fa Fàt-yîm Sṳ-tién*. Nàm-thiên, Taipei, Taiwan, 1996-06. ISBN 957-638-359-5.

[12] William Campbell. *Ē-mn̂g-im Sin Jī-tián*. Tâi-oân Kàu-hoē Kong-pò-siā, Tainan, Taiwan, 1993-06. First published 1913-07.

[13] Bible Society. *Sin-kū-iok ê Sèng-keng (Amoy Romanized Bible)*. Bible Society, Taipei, Taiwan, 1964. Cat. No. AR 065 T, first published 1933.

[14] Chek Hoàn Ko and Pang Tìn Tân. Sin-iok (new testament), the Ko-Tân/Kerymatic colloquial Taiwanese version, 1972-08-31.

[15] Tâi-oân Ki-tok Tiúⁿ-ló Kàu-hoē Im gȧk Úi-oân-hoē, editor. *Sèng-si*. Jîn-kong, Tainan, Taiwan, 1980-03. First published 1953-02.

[16] Bible Society. *kai nua Cemas (Paiwan New Testament & Shorter Old Testament)*. Bible Society, Taipei, Taiwan, 1993. Cat. No. PW63PT-227T.

[17] Iok-hān Hoâi, editor. *ti vuvu katua vatu katua ngiaw (Grandfather, the dog and the cat)*. Bible Society, Taipei, Taiwan, 1987-03.

[18] N.A. Nevskij, Sìhóng Bái, Boris Riftin, and Zhōngchéng (Bāsūyǎ Bóyīzhénǔ) Pǔ. *Táiwān Zhōuzú Yǔdiǎn*. Táiyuán, Taipei, Taiwan, 1993. ISBN 957-9261-41-5.